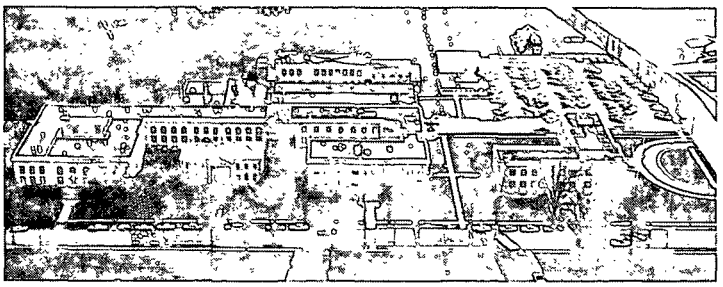

THE INSTITUTE OF PAPER CHEMISTRY, APPLETON, WISCONSIN

IPC TECHNICAL PAPER SERIES
NUMBER 137

# A MODIFIED APPROACH TO COMMUNITY COMPARISON IN AQUATIC SYSTEMS

MICHAEL TESMER, JOHN TEED, AND MICHAEL MISCHUK

JANUARY, 1984

# A MODIFIED APPROACH TO COMMUNITY COMPARISON IN AQUATIC SYSTEMS

Michael Tesmer, John Teed, and Michael Mischuk

## SUMMARY

The study of biological communities, particularly aquatic ecosystems, has been for the past decade an important part of water quality monitoring in this country's streams and lakes. These studies have typically generated large amounts of technical data which needed to be analyzed and presented in some understandable fashion. Because of this, mathematical reductions, such as diversity indices and cluster analysis, were utilized to reduce this data to more simple comparative values. These mathematical techniques have succeeded in data reduction, but more often than not some loss of important information occurred.

The following presentation describes a technique to compare biological communities. The comparison is based upon the structural attributes of the community (population identities and function), expressed as a similarity index.

In order to develop this modified approach, it was first necessary to understand the basic characteristics of clustering. The preferred clustering strategy for ecological data are discussed along with the primary needs of a similarity measure. Likewise, a discussion of the basic attributes of biological communities was necessary in order to define their important properties.

With the above principles understood, the problems associated with the comparison were analyzed. Current indices are designed using either species abundance information or species identity information. The modified approach developed in this paper used cluster analysis and all available population data. It is based on the following ideas: first, that the fundamental requirements of the cluster analysis are met by existing data; second, that these requirements

could be used to empirically derive a similarity index that best fit the analysis; and third, that the requirements of the cluster analysis and derived similarity index could be compared against the ecological principles of the community concept. In order to initiate the approach, it was necessary to answer two basic questions. One, what is our real purpose in performing a cluster analysis, and two, what are the requirements of such an analysis? The answer to the first question is to group the sampling units into clusters that display the levels of natural association between sampling units, at the community level. The answer to the second question is more difficult, since the basic data unit or entity for comparison was to be selected, and a decision was to be made regarding the perspective of the set of entities to be compared. A discussion of the sampling unit as it equates to a fully censured community was necessary in order to answer the second question.

With the above conditions satisfied, a discussion of the index is presented. It is based along a three partition system. The first partition addresses the aspect of relative similarity in the abundance information. The second partition addresses the relative importance of the abundance information, thus handling the aspect of partially present species. The third partition addresses the species identification. The justification for the use of this index as well as statistical assumptions are discussed.

# INTRODUCTION

Simply stated, the biological community concept involves a group of populations living in association with, and in close proximity to, one another. The composition and structure of the community is a function of the environment it inhabits. Any changes which occur in that environment will subsequently be reflected in the indigenous biotic community. These resultant changes in the community can serve as important diagnostic criteria. If properly analyzed, they can help to broaden our understanding of environmental dynamics and serve to highlight the nature of pollutant interferences. As such, the community concept has gained considerable importance as one of the premier units in aquatic ecology (Cairns, 1979).

Community oriented studies typically generate large amounts of raw data. Information concerning the abundance and distribution of many populations is commonly obtained. In order to evaluate such complex data at the community level, it becomes necessary to summarize the information along the guidelines of a preconceived community unit. While a precise definition of the "community unit" continues to evade ecologists, many techniques have been developed for reducing the data prior to final analysis. Two of the more promising techniques appear to be ordination (Culp and Davies, 1980) and cluster analysis (Green and Vascotto, 1978). Recently, a combination of these two techniques has also been used (Flox, 1980, and Gauch and Whittaker, 1981). Unfortunately, these techniques have met with only limited success. The large assortment of testing procedures available has led to some skepticism regarding the ecological validity of the results.

We have found cluster analysis to be particularly useful in displaying the inherent patterns of similarity between biological sampling units. This is consistent with others (Cairns and Kaesler 1969, Day et al. 1971, Levings, 1975, and Eagle, 1975) and appears to be an appropriate use for cluster analysis (Williams, 1971, Clifford and Stephenson, 1975, and Green, 1979). In order to use cluster analysis with some degree of confidence, we found it necessary to review the basic concepts involved. It became quite evident that the results of a cluster analysis depended upon (1) the fundamental qualities of the hypothetical community unit involved, (2) the particular measure of similarity used in data reduction, and (3) the clustering strategy applied to the similarity matrix. The same data could be manipulated to produce conflicting results by merely changing any one of these three characteristics. The available literature was a source of further confusion. A multitude of approaches currently exist and appear to stem from fundamental differences regarding the inherent biological principles which govern both the community unit and its structure.

As a result, we chose to identify the information we desired from a cluster analysis and to develop a procedure which extracted such information accordingly. To accomplish this task, the basic concept of the biotic community was reviewed, and the principle biological features responsible for community structure were identified. The more common similarity indexes were then examined for their ability to represent these structural attributes so that the major difficulties with their application could be identified. Finally, the structural attributes were expressed in a similarity index based on a combination of desirable characteristics found in other similarity indices.

## A BASIS FOR CLUSTER ANALYSIS

Our primary interest in aquatic communities stems from a concern over the changes in community structure which result from pollutant induced stresses. While the impact of a pollutant may vary among community members, the net result of such changes at the community level is of substantial ecological importance. At present, aquatic studies aimed at impact assessment primarily rely upon a taxonomic inventory of the community and the ability to quantitatively identify spatial or temporal variations which exist in the more important species populations within that community. After the biological data have been analyzed by traditional parametric methods, it still remains difficult to evaluate the combined effects of the observed disturbances upon the community. This is due both to the exclusion of data (minor species) not amenable to parametric statistics, and to the extremely complex nature of biological communities. Thus the need exists for either a better means of data collection or for a method of analyzing the net effect of variations in the species populations upon the overall community structure.

Like any research endeavor, restraints in time and money will dictate the quality and amount of available data. Thus, priorities must be set regarding which desirable characteristics the data should possess. The study design commonly employed provides an estimate of total diversity and maximum information about the major species populations. This design is not unintentional. In order to provide sufficient information about all the major species populations, the sampling intensity would become extremely cost restrictive. In addition, more species of even lesser frequency would undoubtedly be found. The necessary cutoff becomes a point where sufficient information about the major species populations has been collected to allow for the statistical verification of any

major changes attributable to the experimental variables under investigation. Elliott (1977) and Green (1979) provide excellent guidelines for both the proper study design and interpretation of results for aquatic studies. We concur fully with the use of such criteria and recommend that such procedures be strictly adhered to.

By virtue of the law of diminishing returns, most investigators will continue to be faced with information gaps in the data for evaluating community structure. Well designed studies do, however, contain sufficient information for estimating major changes in community structure. The use of cluster analysis as a summative technique can be applied to display patterns of similarity existent between sampling units at the community level. The procedure can provide a reasoned subjective assessment of community similarity based on an objective analysis of the data. As such the analysis can only imply relationships between sampling units based upon their common structural attributes. The analysis, therefore, is used as a supplement to the information gained from the parametric statistics, not as a replacement for them. The cluster analysis provides the observer with a tool for examining a greater portion of the collected information, in a reduced form, at the community level.

## CLUSTERING CHARACTERISTICS

There are many options available when selecting an appropriate classificatory method. The actual choice should be dependent upon the nature of the data and the desired information. Our interest is primarily with data sets consisting of randomly collected quantitative sampling units. The desired information concerns the relative degree of similarity between sampling units based upon the species population present. Thus, the sampling units comprise the objects to be compared, and the individual populations or their inherent biological characteristics comprise the attributes for comparison.

Polythetic agglomerative hierarchical clustering methods have been preferred for this type of ecological data (Clifford and Stephenson, 1975, and Sokal and Sneath, 1963). These methods typically involve two independent, yet complementary, operations to produce the final result. First, the multiple attributes are examined for each possible pair of objects and reduced to a single summative value in each paired comparison. All possible pairs of entities are examined, and the results are presented as a matrix of paired comparisons. Second, the matrix of paired comparisons is subjected to an appropriate clustering strategy to group the entities. The objective of the reductive process is to summarize the similarity relationships existent between entities based on the net effect of all the attributes considered. The process is "polythetic" in that it considers more than one attribute/paired comparison, and "agglomerative" in that all attributes are considered simultaneously. The objective of the clustering strategy is to group or "cluster" the entities based on the similarity relationships present in the paired comparisons values. A nested or "hierarchical" clustering strategy builds such groups sequentially,

through the continual merger of smaller groups. The patterns established during the merging process are typically displayed in dendrogram or tree structure form.

Two important decisions remain to be made before this classificatory method can be applied: (1) the actual similarity measure for producing the matrix of values must be selected, and (2) the desired strength of the clustering properties must be chosen. The decisions regarding the clustering properties are basically mechanical in nature. Some clustering schemes tend to be weakly clustering, whereas others are strongly clustering. In order to select an appropriate clustering strategy we turned to the desired information. The ideal clustering strategy would allow each sampling unit to carry equal weight and remain independent during the initial comparison of entities. All possible pairs of entities can then be examined with each entity possessing an equal opportunity to be most similar to any of the remaining entities. Once the two most similar entities have been identified they can be joined, thus reducing the number of entities by one. The remaining entities should then be given an equal opportunity to either join the most similar pair of entities or to start a new group with any remaining entity. In either case, the number of entities will again be reduced by one. This process will be repeated until all entities and groups of entities are joined. The end result will be a progressive clustering of entities and entity groups as similarity decreases. The requirement of "equal opportunity" during the clustering process would allow the degree of clustering to be a characteristic of the entities, rather than one of the clustering method. While no clustering strategy presently maintains equal opportunity to all entities after the initial comparison, the pair-group methods developed by Sokal and Mitchener (1958) appear to be the best available. These clustering methods tend to be intermediate in their ability to form clusters

(Clifford and Stephenson, 1975) and appear to produce the least amount of distortion between the original similarity coefficients and their position in the dendrogram (Sokal and Sneath, 1963). In our applications we preferred the weighted pair-group method which places greater (numerical) weight upon remaining entities than on entities already joined in a group.

The similarity measure selected to compare entities is most critical in determining the outcome of the cluster analysis. It has also proved to be the most controversial for ecological data. Pinkham and Pearson (1974) list 11 similarity indexes in current use as well as introducing another. Clifford and Stephenson (1975) review 5 categories of similarity measures, of which similarity indexes are only one. The large assortment of similarity measures makes the selection of any one such measure extremely difficult. As with the clustering strategy, we again considered the desired information. Our primary need is a similarity measure that will examine multiple attributes between any two independent entities. The measure of similarity must be based on the inherent biological principles governing community structure, rather than artifacts of the mathematics involved. Rather than selecting several similarity measures for testing, we chose to reexamine the community concept from a practical viewpoint. The structural attributes identified were then used as a basis for examining some of the more common similarity indexes.

## THE BIOTIC COMMUNITY

To some degree, all organisms appear to live in association with one another, rather than as independent beings haphazardly strewn about (Odum, 1971). The diverse nature of the environment and of the organisms that inhabit it result in some associations being much stronger than others. The community concept appears to have originated out of the need to categorize these associations. The logical approach was to provide categorical units of workable size based on organism groups displaying strong association or along environment gradients where such strong associations occurred. Differing opinions among ecological theorists (Clifford and Stephenson, 1975) concerning which categorical breakdown was best has resulted in a lack of clear definition surrounding the boundaries and compositions of biological communities. In its broadest sense, the "community" can include all the biological components of the entire ecosphere. In its narrowest sense, it has been equated to the microbial populations inhabiting a sand grain. In its practical application, the community represents a division or subdivision of the ecosphere along the major biotic or abiotic discontinuities which occur in the ecosphere.

Although the existence of these various communities appears to be a natural phenomenon, the actual size and composition of a community appears related to the observer's chosen perspective. Sumich (1976) defines a community as, "an ecologically integrated group, consisting of all the populations living in a given limited area." Odum (1971) and Kendeigh (1974) consider communities to have variable size and composition. Both investigators distinguish between major communities, which are sufficiently large and complex enough to be self supporting, and minor communities, which are much less complete and not self supporting.

Based on the flexibility of these definitions, it would appear that the community is a contrived entity for the benefit of the observer. Its boundaries and composition can vary, but must be defined in some logical manner based on the ecological characteristics of the system under study. While the selected community unit has the appearance of a discrete entity, it is in reality, a portion of a larger system where, over time, interactions will occur. A biotic community will possess two distinct sets of boundaries: (1) those that are selected arbitrarily by the observer, and (2) those that are dictated by the environment. The arbitrary boundaries are fixed by the categorical selection of the organisms to be examined and the habitat(s) to be studied. The environmental boundaries are then fixed by the variety of organisms and habitats confined by these arbitrary boundaries but remain subject to temporal fluctuations due to interactions with the surrounding ecosystem.

## COMMUNITY STRUCTURE

Regardless of the conceptual differences that exist between variously defined communities, all biological communities share similar structural attributes. The species and their inherent associations are the common foundational link between communities. While the ultimate structural complexity of a community is dependent upon its defined boundaries, the individual attributes contributing to the overall structure are not.

The individual organism comprises the smallest biological unit of interest in a community. However, its dependence upon the success of all its species' members for the continuation of its population makes the species population the basic unit of ecological interest to the community. The relative importance of a species population as a unit will be dependent upon some characteristic of the organisms of which it is composed. At any given point in time, a biological community will be composed of a finite set of such species populations where: S = (a,b,c, . . . n). Each species possesses two important structural characteristics: (1) its identity (i) which is universally constant, and (2) its functional importance in the community. The first characteristic is directly determinable and contributes to the size of the finite set of species (S). The second characteristic is not directly determinable and must be estimated from some measured variable, such as numerical abundance ($X_i$), or biomass ($B_i$). The relative importance of a species population to a community can be very different depending upon the chosen measured variable. In the remainder of this discussion we will be referring to numerical abundance as the measured variable. The principles, however, apply to any measure used.

In a typical community, S will be composed of a few relatively abundant species, with progressively more species of lesser abundance (Odum, 1971, Pielou, 1977). The abundant species are typically viewed as being dominant or of greatest importance within the community, whereas the infrequent and rare species are of lesser importance to the community. The distributional pattern established by the proportionment of abundance among the species is best illustrated in a species abundance curve (Fig. 1).

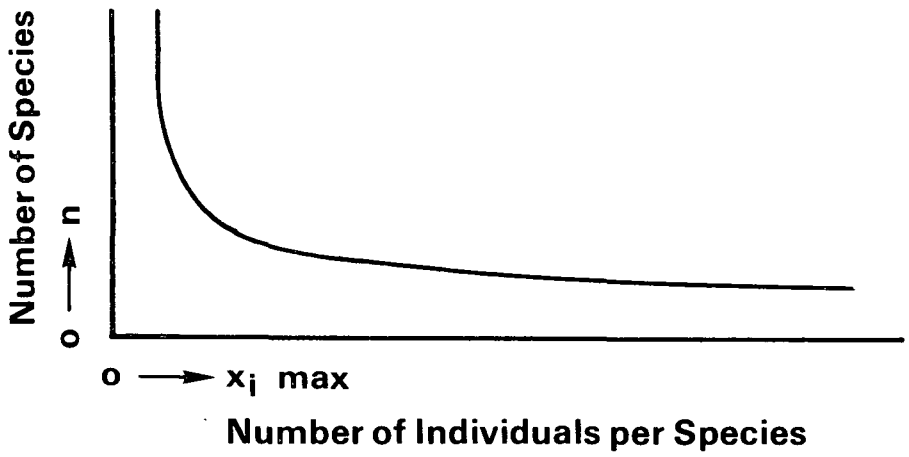


Figure 1. A hypothetical species abundance curve.

The shape of the species abundance curve is dependent upon the distribution of the different species frequencies and may fit any one of several common mathematical distributions (Pielou, 1977). Both the number of species present and the degree of dominance of the major taxa are represented. This aspect of community structure has been rigorously examined (Pielou, 1977) and is the basis for such comparative techniques as diversity indices and analyses involving the fitted mathematical distributions. Unfortunately, these comparative techniques do not consider species identity. While each species is represented, its identity remains anonymous and cannot be located in the curve.

As a result these techniques are of only limited value in community comparison. In order to adequately represent community structure, the identity of each species must be represented as well.

## PROBLEMS OF COMPARISON

While the community unit appears to be valid and the attributes of community structure are easy to identify, it has proven difficult to adequately represent them in comparative analyses. Much of the difficulty stems from (1) the controversy surrounding the community perspective, (2) the nature of the typical data base, and (3) the nature of the attributes themselves.

It is common practice to treat the community unit as being bounded by either (1) the entire study area, (2) the major strata within the study area, or (3) the spatial areas of interest (control vs. experimental) within the study area. A set of random samples can then be collected from within the selected community units with the number of individuals in each species/sampling unit being the measured variable. Random samples are required to insure the independence of items from extraneous and experimentally uncontrolled sources of variation (error). The large community unit provides sufficient sampling area so that the distribution patterns of the major species can be determined. Each of these distributions can be statistically evaluated to identify significant spatial differences which exist either within a community unit or between community units. Unfortunately, the accompanying changes in the minor species cannot be determined due to the lack of sufficient information about their distributions.

When a similar rationale is attempted at the community level, several critical problems arise. If the community unit is bounded by the study area or some large physical portion of the study area, the principal comparisons between sampling units will involve intracommunity variations. Because the parent community is too large to be fully censused, the community parameters of S and each

The traditional approach has been to control as many of these problems as possible rather than to reexamine the community unit. The wide variety of similarity indexes primarily results from the different approaches used. It is beyond the scope of this paper to discuss all of the available indexes and their associated problems. An excellent description of the more common and widely used indexes can be found in Clifford and Stephenson (1975) and Hellawell (1978).

Early investigators such as Jaccard (1908), and Czekanowski (1913), as well as many others (see Clifford and Stephenson, 1975), relied strongly upon the binary aspects of the data (presence/absence) and completely ignored the problems associated with the abundance estimates. All of these approaches are initially based upon the Simple Matching Coefficient (SMC) and differ primarily in the weighting of characters contained in it. A 2 x 2 table is constructed for each pair of entities:

| | | Entity 2: Present | Entity 2: Absent |
|---|---|---|---|
| Entity 1 | Present | a | b |
| | Absent | c | d |

where

a = the number of species mutually present in entities 1 and 2

b = the number of species present in entity 1 only

c = the number of species present in entity 2 only

d = the number of species mutually absent in entities 1 and 2.

and

$$SMC = \frac{a + d}{a + b + c + d}$$

Similarity scale

Low ——> High
0 ——> 1

Mutually present species (a) and partially present species (b and c) are typically tallied. Jaccard (1908) assigned equal value to each species and ignored mutual absences:

$$\text{Jaccard} = \frac{a}{a + b + c} \qquad 0 \longrightarrow 1$$

Czekanowski (1913) also ignored mutual absences but assigned double weight to mutually present species, thus considering them twice as important in the comparison:

$$\text{Czekanowski} = \frac{2a}{2a + b + c} \qquad 0 \longrightarrow 1$$

Some researchers chose not to ignore the mutually absent species and even considered such species equal to mutually present species worthy of additional weight (Sokal and Sneath, 1963):

$$\text{Sokal and Sneath} = \frac{2(a + d)}{2(a + d) + b + c} \qquad 0 \longrightarrow 1$$

All of these similarity indexes treat the data qualitatively and are heavily weighted toward the structural attributes of species identity and species occurrence. Because information is excluded, such indexes underestimate the contribution of dominant species and overestimate the contribution of lesser species.

Attempts to better integrate the community concept with the similarity index necessitated the incorporation of the species abundance information. The simplest approach was to modify the qualitative indexes by summing the species abundances in place of the species occurrences within the characters of the 2 x

2 table. This technique is of limited value because it provides for no comparison between entities for the mutually present species. Further modifications were then applied. Hellawell (1978) describes such an extension of Czekanowski's (1913) coefficient:

$$\text{Modified Czekanowski } \frac{2W}{A + B} \qquad 0 \longrightarrow 1$$

where

$$W = \sum_{i=1}^{n} \min (X_{i1}, X_{i2})_a$$ (the sum of the minimum abundances of mutually present species in entities 1 and 2).

$$A = \sum_{i=1}^{n} \min (X_{i1},)_a + \sum_{i=1}^{n} \max (X_{i1},)_a + \sum_{i=1}^{n} (X_{i1}, 0)_b$$ (the total abundance in entity 1).

$$B = \sum_{i=1}^{n} \min (X_{i2})_a + \sum_{i=1}^{n} \max (x_{i2})_a + \sum_{i=1}^{n} (0, X_{i2})_c$$ (the total abundance in entity 2).

Another modified approach known as the Bray-Curtis (1957) measure has been referred to as the complement of Czekanowski's (1913) coefficient (Clifford and Stephenson, 1975):

$$\text{Bray-Curtis measure} = \frac{\sum_{i=1}^{n} | X_{i1} - X_{i2} |}{\sum_{i=1}^{n} (X_{i1} + X_{i2})} \qquad 1 \longrightarrow 0$$

The Bray-Curtis measure differs from the previous indexes in that it represents a measure of dissimilarity. The scale is therefore reversed with 0 being most similar and 1 being least similar.

In these earlier indexes the abundance information was summed across the various species to determine the final ratio or coefficient. Mutually absent species were automatically ignored because they provided no numerical input to the summations. The final coefficient reflected the combined differences between the species present in the two entities, and was easily influenced by extremely abundant species. Although this has been typically identified as a problem, it can be argued that the similarity between two entities, strongly dominated by only one or two species, should be primarily a function of those species. The actual problem is whether the "weighting" effect is too dramatic when it results in the virtual elimination of the information from the lesser species.

Lance and Williams (1966) attempted to overcome this problem by modifying the Bray-Curtis measure so that the contribution of each species could be evaluated separately.

$$\text{Lance-Williams} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{|X_{i1} - X_{i2}|}{(X_{i1} + X_{i2})} \qquad 1 \longrightarrow 0$$

The dissimilarity displayed by each species was evaluated separately. The results from each species comparison were then summed and divided by the number of participating species. Like the Bray-Curtis index, the scale was reversed and represented a dissimilarity measure.

Pinkham and Pearson (1974) also developed an index based upon the sum of the intraspecies contribution to similarity:

$$\text{Pinkham-Pearson} = \frac{1}{n} \sum_{i=1}^{n} \frac{\min (X_{i1}, x_{i2})}{\max (X_{i1}, x_{i2})} \qquad 0 \longrightarrow 1$$

Unlike the Lance-Williams measure, Pinkham and Pearson relied upon the direct ratio of abundance for each species, resulting in a return to the more traditional similarity scale of 0 (low) to 1 (high).

Several additional problems develop when the similarity indexes rely upon individual species evaluations. First of all, the use of ratios on meristic variables can result in substantial inaccuracy (Sokal and Rohlf, 1969). In an index such as Lance-Williams or Pinkham-Pearson, where a series of ratios are summed, this problem has the potential of being compounded many times. It can be especially troublesome when these indexes are applied to qualitative data or when the number of minor species in a community is high. Secondly, a species that is only partially present ($X_i$, 0) in the comparison will always produce a ratio equivalent to the minimum similarity, regardless of the size of the measured population. Thus, two species whose populations are (100, 0) and (1, 0) will be considered equally dissimilar. Thirdly, the ratio produced by mutually absent species (0, 0) is mathematically indeterminable. Such species must either be ignored or assigned a value.

The problems associated with partial or complete absenteeism have been approached in several different ways. Some investigators (Lambert and Dale, 1964, Lance and Williams, 1967, and Stephenson et al., 1972) chose to eliminate the rare species, thus removing the majority of partial and complete absence matches. The remaining mutual absences (0, 0) were either assigned a value (% = 0 or 1) or ignored. The remaining partial absences were treated by assigning a small positive integer to the absent value (0 = 0.01), which produced a varied ratio better reflecting the degree of difference between partially present species. Pinkham and Pearson (1974) chose to include all minor species in the data set, but assigned a value to both the partial and complete absence matches. In fully

quantitative data they considered the absence of an organism to be a real event and assigned significance. Mutual absences (0, 0) were assigned a value equal to maximum similarity. The partial absences ($X_i$, 0) were assigned a value equal to minimum similarity, regardless of the population abundance ($X_i$) present at one of the sites being compared.

The similarity indexes designed to incorporate the abundance information are biased toward the abundance relationships. While species identity is accounted for in the individual species comparisons, its input to the numerical result is negligible. Thus, these indexes represent the opposite extreme from the early indexes and are heavily weighted toward only the abundance relationships.

All of the similarity indexes presented and their associated manipulative techniques are difficult to justify from the biological point of view. It must be remembered that each species possesses two important attributes with regard to community structure: its identity and its abundance. The reliance upon one or the other attribute cannot provide for a good, flexible, comparative analysis capable of functioning over the extremes in community structure. For example, the elimination of rare species is based on the assumption that the majority of information is contained in the frequently encountered abundant species. While this may be true in communities where species diversity is low, it does not hold true for highly diversified communities where no apparent species clearly dominate. On the other hand, the assignment of value to species that are partially or mutually absent can lead to an overemphasis of their importance much like the early qualitative indexes. In data sets possessing many rare species or where the species distributions are highly uneven, the partial and complete absence matches can dominate the similarity index calculations.

Because the value of 0 in a quantitative ratio is not comparable, the zero in a match has no utility whether the match is partial ($X_i$, 0) or complete (0, 0). The assignment of either maximum or minimum value based on the presence of a positive integer ($X_i$) is purely arbitrary, and the values of 0 become inconsistent over the sample set. Not only are the values of 0 inconsistent, but the information contained in the partially present matches (where $X_{i1}$ or $X_{i2} > 0$) will also be lost unless examined in such a way as to eliminate the problematic zeros.

To review briefly, the available similarity indexes appear to suffer from three recurrent problems.

1. The indexes are designed around either the species abundance information or the species identity information which causes a disproportionate emphasis on one or the other attribute.

2. The mathematical formulation is rigid or inflexible resulting in poor applicability over the extremes in community structure.

3. The partial and complete absence matches are either manipulated in some manner (partial elimination - arbitrary assignment of value) or ignored.

## A MODIFIED APPROACH

There appears to be no easy solution to the problems confronting the use of cluster analysis as a summative tool in community comparisons. The most useful data will continue to be designed around the requirements of the parametric statistics used to analyze individual species populations rather than entire communities. Community structure will, of necessity, remain as a secondary issue. This attitude is precipitated from the vague and often unruly nature of the community unit. Its mathematical reduction is strongly dependent upon which conceptual ideals have been selected to govern the community unit. Different ideals will result in different similarity measures, which, when finally clustered, will result in different clustering strategies for the same data base. Thus, as long as the theoretical aspects of the community concept remain controversial, the statistical analyses needed to examine the community must be fit in retrospect.

Herein, lies the heart of the problem. It is not with the cluster analysis itself, but with the manner in which it has been applied. All too often, cluster analysis, like other multivariate analyses, are applied as an afterthought in an attempt to explain the hitherto unexplainable. There is little value in attempting to "retrofit" any analysis to a set of ideals or the generated data that were established along preconceived guidelines for another purpose. The fundamental ideals must be firmly established so that an appropriate analytical approach can be selected. The data should then be collected along the guidelines of the intended analysis. Where more than one analysis is involved, the data should at least be scrutinized to see if they meet the necessary requirements of the additional analyses as well.

Because the fundamental ideals of the community concept are not firmly established, our approach was to focus on the cluster analysis and the available data. If the fundamental requirements of the cluster analysis were met by the existing data, at least its use could be justified. Furthermore, these requirements could also be used to empirically derive a similarity index that best fit the analysis. Finally, the requirements of the cluster analysis and the derived similarity index could be compared against the ecological principles of the community concept. A decision regarding its applicability could then be made.

In order to initiate this approach we first sought the answers to two basic questions: (1) what is our real purpose in performing a cluster analysis, and (2) what are the requirements for such an analysis?

In biological studies aimed at impact assessment, the real purpose of cluster analysis is to group the sampling units into clusters that display the levels of natural association between sampling units, at the community level. The key to the performance of the cluster analysis lies in the observer's interpretation of the phrases "natural association" and "community level." For our purpose, "natural association" has a dual meaning. Mathematically, "natural association" is interpreted as the level of similarity displayed by the attributes contained in the entities being compared. Ecologically, "natural association" is interpreted as the level of similarity existing between sampling units (entities) based on a comparison of the species populations (attributes) present that share a natural ecological association as well (i.e., all periphyton species, all phytoplankton species, all macroinvertebrate species). The "community level" is interpreted as the level at which all of the species populations belonging to the selected natural association within a sampling unit are examined simultaneously.

The statistical requirements for a cluster analysis are highly flexible and dependent upon its intended use. This results primarily from cluster analysis typically being used to generate hypotheses rather than for testing hypotheses (Anderberg, 1975). In order to perform a cluster analysis it is first necessary to select the basic data unit or entity for the comparison. Next a decision is required regarding the actual perspective of the set of entities to be compared. There are basically two choices: (1) each entity is in itself complete and therefore the true object of interest, or (2) each entity is incomplete and represents an estimate of a much larger population which is the true object of interest.

In the first case, the intended purpose of the analysis is to produce a classification scheme for only the entities contained within the data set. The underlying assumptions are:

1. Each entity is complete.
2. Each entity represents a true object of interest.

The restrictions are:

1. The results are specific for the set of entities involved.
2. Entities from outside the original sample set cannot be compared.

In the second case, the intended purpose of the analysis is still to produce a classification scheme for the entities contained in the data set, but the comparison is based upon the true object of interest: the parent population. Here the assumptions and restrictions are similar to those required for the more familiar parametric statistics.

From a viewpoint of cluster analysis, either of the two cases is a valid approach. The choice is left to the researcher and will be dependent upon the purpose of the analysis. As already identified, our intended purpose is quite clear. First of all, we wish to compare a finite set of sampling units against one another at the community level rather than against a hypothetical "parent community." Secondly, we wish to consider each sampling unit as a true object of interest. Finally, we wish to include all of the species information contained in each entity including those species that do not meet the parametric test criteria. It is evident that the underlying assumptions and restrictions of the first case closely parallel our desired purpose. Rather than attempting to correct for the necessary violations of the second case, as has been done in the past, we can alter the perspective of the data so that the requirements of the first case can be met.

This brings us to a very critical point, that being the data base. Ideally, it would be desirable to collect a data base designed around the requirements of the cluster analysis. This is unlikely because of the need to retain the original data base. The additional cost of a second data set would also be restrictive. Thus the original dilemma: Do we retrofit the cluster analysis to the data base, or do we retrofit the data base to the cluster analysis? Because we feel this issue is central to the problems confronting cluster analysis and critical to the modification we are about to propose, it warrants careful consideration.

As stated previously, the typical data base is designed around the requirements of parametric statistics in order to identify changes in the major species populations. Thus, the majority of the attributes (species) for the cluster analysis can be adequately analyzed under the parametric criteria.

Because of the apparent similarity in the requirements for the generation of the data base and the second case for cluster analysis, one would expect the data base to be highly compatible. In fact, the second case would appear to represent a logical extension of the principles of parametric statistics from that of the species level to that of the community level. However, the parent population in the cluster analysis is actually the parent community. The ability to satisfy the statistical requirements at the community level are extremely difficult. Each participating species (attribute) must first meet these requirements. Because this is virtually impossible, due to the rare species, violations of the underlying assumptions will occur. It then becomes necessary to correct for such violations by adjusting the similarity index, which in effect undermines the foundational aspects of the cluster analysis. Thus, the cluster analysis is made to fit the shortcomings of the data.

While the data base cannot be considered optimal for the first case in cluster analysis either, it does appear applicable provided that the consequences are fully understood. Rather than modifying the cluster analysis to fit the data, the perspective of the data must be modified to fit the requirements of the cluster analysis. The underlying assumptions of the first case require that each sampling unit represent a complete entity and a true object of interest. The sampling unit is, therefore, equated to a censused community. The set of species (S) is known, as is the abundance ($X_i$) of each species contained in S. What was originally perceived as a species population estimate ($X_i$) in a sample, is now a measured population parameter ($X_i$) within a censused community. Only the species populations present in a sampling unit are viewed as members of the censused community. The consequences of the altered perspective are quite dramatic. All sources of error in the species estimates and all sources of natural

variability in the species distributions are removed from the analysis and held as extrinsic attributes. Only the intrinsic data enter into the analysis. Because the results of the cluster analysis are specific for the set of entities, the selection of those entities is critical to the outcome. The random selection of the sampling units may cause the exclusion of an important entity.

While the perspective of the data can be manipulated to meet the requirements of the first case, it is done so at considerable risk. The data is made more compatible with the intended purpose of cluster analysis, but places the additional burden of identifying the sources of variation on the observer. If we are willing to accept such a risk, the utility of the cluster analysis is greatly broadened. Any set of sampling units can be compared. The observed classification scheme can then be compared against any or all measured sources of variation to determine which factors appear most influential.

Let us for the moment then, change our perspective of the data so that the requirements of the first case are met. The community perspective is changed from that of a single large community whose population parameters are estimated, to that of a series of much smaller communities, each of which is fully censused. Because each community is in itself a complete entity and the set of species may vary between entities, it will be necessary to compare all possible community pairs separately. In this manner, equal opportunity of being most similar to any entity is afforded to each community. Only the attributes relevant to each paired comparison need be considered. Therefore, each paired comparison can be based upon the similarities that exist between the two sets of population parameters where:

$S_a$ = the species present in community a

$S_b$ = the species present in community b

$T_{ab}$ = the species present in either or both communities a and b

$K_{ab}$ = the species mutually present in communities a and b

$X_{ia}$ = the abundance of the ith species in community a

$X_{ib}$ = the abundance of the ith species in community b

$D_{ab}$ = the total abundance of community a + community b

The ultimate goal in each paired comparison is to produce a single "Coefficient of Similarity" that numerically reflects the degree of association existing between the two communities. This is accomplished by comparing the attributes of the two entities via a similarity index. The index must be formulated in such a way that it compares the entity attributes based on the conceptual nature of the community unit and its inherent structure. Because the community unit was selected to meet the requirements of the cluster analysis, we can now develop an index specifically suited for the cluster analysis.

By developing the similarity index empirically, we can take advantage of the modified community perspective and greatly reduce or eliminate many of the problems that plagued earlier indexes. In each paired comparison, the set of attributes will be composed of the combined set of species ($T_{ab}$) for the paired communities. As a result, only two possibilities exist for each attribute: (1) mutual presence ($X_{ia}$, $X_{ib}$) or (2) partial presence ($X_{ia}$, 0), (0, $X_{ib}$). The mutual absence matches (0, 0), so difficult to control previously, are eliminated. Because $X_{ia}$ and $X_{ib}$ are known parameters, mutually present species abundances can be compared by direct ratio, greatly reducing the mathematics involved. Only the problematic zeros in the partial presence matches remain as a major stumbling block.

Another advantage to the empirical derivation of the index is that we can build in certain additional measures aimed at improving the comparative process. One of the major drawbacks in community comparison has been the inflexibility of the similarity indexes. Either they focus on species identity information or species abundance information. The inherent structural characteristics of the modified community unit are the same as for any community. Each species or community attribute possesses the structural characteristics of identity and functional importance (abundance in this case). With the population parameters determined in the censused communities, it should be possible to partition these attribute characteristics during the comparative process. If each partition is of equal scalor value and all partitions are formulated from a common perspective of similarity, they can ultimately be combined in the index to produce the desired coefficient. The extremes in community structure can then be adequately represented.

The manner in which the characteristics are partitioned is dependent upon the relative importance of each characteristic and its nature with regard to the aspects of similarity. It is generally accepted that the characteristic of functional importance contains the majority of the similarity information. The nature of functional importance with regard to similarity is quite complex and dependent upon two principal relationships: First, the functional importance of a species is related to its contribution to the total abundance of the community; the greater the contribution the greater its importance. Second, the degree of similarity displayed by a species, between two communities, is related to the numerical nearness of the abundances for that species; the nearer the abundances are to equality the more similar is the species. Thus, both the similarity of the individual species and the relative importance of each species

must be considered. While the characteristic of identity is regarded as somewhat less important than the abundance information, its contribution to similarity is easily lost in the quantitative indexes. Its importance lies in the fact that identity displays the degree to which the paired communities share similar species. The ideal situation, then, would appear to be an index consisting of three partitions. The first two partitions would address the aspects of functional importance, whereas the third partition would address the aspect of identity. If each partition is of equal weight in the index, the abundance information will account for 2/3 of the similarity measure and the identity information 1/3.

The existence of partial absence matches ($X_i$, 0) in the data base make it difficult to completely partition either the relative importance or the relative similarity for each species. A partially absent species theoretically carries only half of the necessary abundance information to determine either aspect of functional importance. In order to accommodate the desire to partition or at least consider both relative importance and relative similarity the partitioning must be done indirectly. The distribution of abundance between each of the mutually present species ($X_{ia}$, $X_{ib}$) can be recognized as a positive aspect of similarity. The relative similarity of each species can be judged as can the combined contribution of each species to the total abundance of the paired communities. The partially present species, however, must be recognized as a negative aspect of similarity. With only half the necessary information present, the difference in species abundances reflects the degree to which each partially present species is dissimilar. For example, a species whose abundance is (0, 1) can be seen as being dissimilar to the point that it is just present in one community and completely absent from the other. This situation represents the maximum level of similarity a partially present species can possess.

Clearly, the species whose abundance is (0, 100) is even more dissimilar than the first. In order for the information contained in the partially present species to be useful, it must be examined in such a way as to display its reductive effect upon similarity rather than its additive effect upon dissimilarity. The use of individual abundance ratios, the assignment of maximum dissimilarity ($0/X_i = 0$) to such ratios, and the assignment of a small position value to each zero ($0.01/X_i$) within such ratios are therefore unacceptable. The inclusion of any such ratios will tend to overemphasize the importance of the partially present species.

The key to partitioning the abundance relationships lies in the ability to perceive both the mutually present species and the partially present species from a positive aspect of similarity. The abundance information in the mutually present species will constitute the basis for the calculation of both relative similarity and relative importance. The negative effect of the partially present species will then have to be detracted from the measures. Because they contain only half the abundance information of the mutually present species, their negative impact will be considered only once.

If we follow this approach we can now consider the relative similarity of the mutually present species independent of the partially present species. The ratio of the minimum abundance to the maximum abundance

$$\left(\frac{\min X_{ia}, X_{ib}}{\max X_{ia}, X_{ib}}\right) \qquad (1)$$

can be calculated for each mutually present species. Complete similarity for each species is equated to unity, and similarity is reduced as the difference between maximum and minimum abundance grows. If each mutually present species

is so treated, the results of each ratio can be summed and the sum divided by the number of participating species.

$$\frac{1}{K_{ab}} \cdot \left[ \sum_{i=1}^{K_{ab}} \frac{\min (X_{ia}, X_{ib})}{\max (X_{ia}, X_{ib})} \right] = SV_1 \qquad (2)$$

Equation (2) constitutes the first partition and represents the aspect of relative similarity in the abundance information. It is similar to Pinkham and Pearson's equation in that it uses min/max ratios. However, the partial and complete absence matches have been removed. The similarity value ($SV_1$) will range from $0 < SV_1 \leq 1$. While species identity is recognized, it does not influence the outcome because the number of participating ratios is equal and divided by the number of participating species. Any species can participate in the equation provided it meets the requirement of being mutually present. All species entering the equation carry equal weight in the outcome.

The aspect of relative importance is considerably more difficult to perceive than that of relative similarity. The relative importance of a species can be perceived as either a function of each species, or as a cumulative function of similarity. In order to keep the comparisons of relative similarity and relative importance separate, and to remain consistent within the index, we chose the latter case. The total abundance of the paired communities was assumed to represent the maximum obtainable abundance if all species contributed positively to similarity. Each species could then be given equal opportunity to contribute to, or detract from, that value. The relative importance of each species would then be based on the size of its numerical contribution or reduction. Mutually present species, being a positive aspect of similarity, were considered to represent a positive contribution. Partially present species,

being a negative aspect of similarity, were considered to represent a negative contribution. This in effect is a ratio of the sum total of the mutually present species to the total abundance of the paired communities:

$$\frac{1}{D_{ab}} \cdot \left[ \sum_{i=1}^{K_{ab}} (X_{ia} + X_{ib}) \right] = SV_2 \qquad (3)$$

The similarity value ($SV_2$) will again range from $0 < SV_2 \leq 1$ and decreases as the cumulative abundance of the partially present species increases. This equation is similar to the modified Czekanowski index except that the total abundance of mutually present species is used rather than twice the sum of the minimum abundances of mutually present species.

While both of the partitions using the abundance information rely upon species identity, neither assigns any numerical significance to it. Thus the basic attribute characteristics of identity and abundance have been separated. To include the influence of species identity, the third partition in the index must be included. Because the importance of identity is related to the way in which the species are shared, we considered each species to carry equal weight. To satisfy this requirement, a simple ratio between the number of mutually shared species to the total number of species was used:

$$\left(\frac{K_{ab}}{T_{ab}}\right) = SV_3 \qquad (4)$$

The similarity value ($SV_3$) will again range from $0 < SV_3 \leq 1$. Each species is counted only once, even if it is present in both communities. This provides equal weight to all species in the identity partition. This equation is identical to the Jaccard index.

If the three similarity values are additively combined and divided by three (3), the resultant coefficient of similarity will fall within the range of $0 < C_s \leq 1$. In the event that no shared species exist between the two communities, the coefficient of similarity is hereby defined as being equal to zero. The scale for the modified index then becomes, $0 \leq C_s \leq 1$ where, 0 is equal to minimum similarity and 1 is equal to maximum similarity. The combined partitions are:

$$S = \frac{\frac{1}{K_{ab}}\left[\sum_{i=1}^{K_{ab}} \frac{\text{min. } (X_{ia}, X_{ia})}{\text{max. } (X_{ia}, X_{ib})}\right] + \frac{1}{D_{ab}}\left[\sum_{i=1}^{K_{ab}} (X_{ia} + X_{ib})\right] + \frac{K_{ab}}{T_{ab}}}{3}$$

## JUSTIFYING THE APPROACH

Actually, this entire approach is not new but represents a refinement of existing techniques. First of all, the cluster analysis was never intended to replace parametric statistics. It tended to mimic these statistics only because the typical data bases being investigated had been designed to meet the requirements for the application of parametric statistics. Many of the attempts by researchers to control the problems afforded cluster analysis by such mimicry inadvertently changed the community perspective similarly to what we have done. For example, in choosing to ignore the mutual absence matches (Jaccard and Czekanowski), the sampling units were effectively kept from influencing one another. This is the same effect that occurs in assuming the sampling units to be complete entities. In the quantitative indexes of Bray and Curtis, and Lance and Williams, the properties of the parent species populations are ignored. This effectively equates an attribute estimate to that of a known population parameter. The importance of the approach, however, is not that it has been tried before. It is the fact that the approach is now structured around the requirements of the cluster analysis, rather than the nature of the data base.

In order to meet the underlying assumptions of the cluster analysis the community perspective was changed from an infinitely large community to small censused communities. By changing the perspective the sampling unit is merely replaced by the censused community. This is identical to the approach commonly used in diversity index measures (Pielou, 1977) and has no effect upon the data set other than how it is perceived. The altered community perspective is actually quite compatible with current ecological theory. As discussed earlier, the theoretical basis for the community is vague and open ended. If we accept the existence of communities in the first place, we must also be willing to

accept the fact that they can be of different sizes. The community perspective can be from any vantage point chosen by the observer, provided its boundaries can be sufficiently defined. For our purpose then, we must define the community unit as the assemblage of organisms under study (the selected natural association) inhabiting the sampled medium that is contained in a sample unit. In comparison with the more widely accepted definitions, only slight differences exist. First, the community is made smaller. Second, the natural biological associations used to differentiate between the major communities remain intact, except that the number of species populations per community unit may be reduced. Third, the physical boundaries differ in that they do not necessarily reflect naturally occurring discontinuities; however, the physical boundaries are specific rather than left to speculation.

The major impacts of the altered community perspective are in the ability to discount the mutual absence matches and the manner in which partial absence matches can be handled. Under the restrictions of the large community perspective, both the mutual absence of a species and the partial absence of a species had to be treated as real information (all zeros represent real numbers). Although they were treated as real, a considerable difference of opinion concerning their importance remained. The origin of the controversy appears to stem from two conflicting arguments: (1) the majority of mutual and partial absence matches are the result of the low frequency of occurrence in the rare species rather than the ecological principles governing community similarity. Hence, they should be ignored, and (2) the mutual or partial absence of a frequent and abundant species from any two sites is ecologically important. Hence, they should not be ignored. Both arguments appear valid and must be considered. We contend that the modified approach can account for these factors. It must first

be remembered that the similarity index and the clustering strategy are not independent functions and do interact. The modified index in combination with the clustering strategy takes into consideration the partial and complete absence of a species without the need to recognize the zero value as real information or the need to assign them additional value. The similarity index uses only the abundance information in each censused community ($X_i > 0$). Each species is therefore afforded equal opportunity to contribute where it is present. Because the clustering strategy examines all possible paired comparisons, the overall contribution of species will be a function of both the number of sample units it is contained in and its contribution to the total abundance of those sample units. A frequent and abundant species will, therefore, contribute in a majority of the paired comparisons, and its abundance will be important. An infrequent species will only contribute in a few comparisons. Where its abundance is high it will be important; where its abundance is low its importance will be negligible. Neither species will contribute to similarity in those comparisons where they are mutually absent. The absence of a species, whether it be partial or complete, is felt by its inability to contribute to the similarity in those paired comparisons where the absenteeism occurred. In addition, the absence of a frequent and abundant species is felt more strongly than the rare species by virtue of its greater participation in the remaining comparisons.

The elimination of absence characteristics from the comparisons also appears to be reasonable from an ecological point of view. The biological impact of an infrequent and rare species upon the more frequent and abundant species (of similar ecological importance) is of little consequence. Its presence may be considered important because it adds to the stability of the community and because of a potentially important future role, should it become an

ecologically favored species. Its ability, as a rare species, to influence the structural nature of the surrounding community is limited unless its abundance is high. It is even less likely to directly affect experimental units where it is not even found. The biological impact of a frequent and abundant species, however, can be profound. It can directly influence or even control the remaining species populations. Its absence at a site is important in that it can no longer exert its influence. The ecological significance of such an absence is not so much in the fact that it is absent, but in the adjustments the remaining species populations have made in its absence. The importance of a species then, appears to be primarily in the fact that it is present and to what extent, rather than its absence. This is the basis for the modified index as well.

The statistical justification of the modified approach is rather difficult. The assumption that a sampling unit can be considered a fully censused community is not without criticism (Clifford and Stephenson, 1975). Although the population parameters in the censused community are, by definition, known, it is also known (Elliott, 1971) that species populations are subject to considerable variation within their environmental boundaries. Different species occurring in the same environmental boundaries may also display different distribution patterns. In addition, the observed species abundances in each censused community are directly dependent upon sample size. While the community parameters are not subject to sampling error, the comparisons of such communities definitely are subject to such error. No estimates of these sources of error are considered during the process of comparison.

The rationale for accepting this approach to cluster analysis remains with the fact that we wish to base the comparisons on only the intrinsic data. The resultant classification scheme will then reflect the similarity patterns existent in and exclusively for the objects of interest. Only after this relationship is established can we search for the underlying source or sources responsible for the established relationships. The major criticism with such a rationale is "that the methods of analysis are being selected on a subjective basis to prove what it is the observer wants to prove anyway," (Clifford and Stephenson, 1975). Clifford and Stephenson admit that this is probably true to some degree, but not necessarily bad. We also agree and offer the following quote from Ostle and Mensing (1975).

> "Statistics (as a science) deals with the development and application of methods and techniques for the collection, tabulation, analysis, and interpretation of data so that the uncertainty of conclusions based upon the data may be evaluated by means of the mathematics of probability. However, it should be also evident that there is something more to statistics than the routine analysis of data using standard techniques. For example, the reader should realize that the analyses are exact only if all the underlying assumptions are satisfied. Since this is rarely true, much depends on the skill of the researcher in selecting the methods of analysis that best fits the circumstances of the experimental situation being studied."

Thus, the selection of any statistical method is in itself a subjective choice. The important characteristic is that the method selected be objective in its treatment of the data. While the modified approach was subjectively selected

and subjectively structured around the underlying assumptions of the cluster analysis, it was done so in an attempt to make the overall analysis more objective in its treatment of the data.

## LITERATURE CITED

Anderberg, M. R. 1973. Cluster analysis for applications. Academic Press, New York. 359 p.

Cairns, J., Jr. 1979. Biological monitoring - concept and scope in statistical ecology. Vol. II, p. 3-21. J. Cairns, Jr., G. P. Patil, and W. E. Waters, editors. International Cooperative Publishing House, Maryland, U.S.A.

Cairns, J., Jr. and R. L. Kaesler. 1969. Cluster analysis of Potomac River survey stations based on protozoan presence-absence data. Hydrobiologia 34(3-4):414-32.

Clifford, H. T. and W. Stephenson. 1975. An introduction to numerical classification. Academic Press, New York. 229 p.

Day, J. H., J. G. Field, and M. P. Montgomery. 1971. The use of numerical methods to determine the distribution of the benthic fauna across the continental shelf of North Carolina. J. Animal Ecology 40:93-125.

Eagle, R. A. 1975. Natural fluctuations in a soft bottom benthic community. J. Mar. Biol. Assoc. U.K. 55:865-78.

Flos, J. 1980. Ordination and cluster analysis applied to oceanographical data. Estuarine and Coastal Marine Science II:393-406.

Gauch, H. G., Jr. and R. H. Whittaker. 1981. Hierarchical classification of community data. J. of Ecology 69:537-57.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. John Wiley & Sons, New York. 257 p.

Kendeigh, S. C. 1974. Ecology with special reference to animals and man. Prentice-Hall, Inc., Englewood Cliffs, NJ. 474 p.

Lance, G. N. and W. T. Williams. 1967. Mixed-data classificatory programs. I. Agglomerative systems. Aust. Comput. Journal 1:15-20.

Levings, C. D. 1975. Analysis of temporal variation in the structure of a shallow-water benthic community in Nova Scotia. Int. Revue Ges. Hydrobiol. 60(4):449-70.

Odum, E. P. 1971. Fundamentals of ecology. Third edition, W. B. Saunders Company, Philadelphia. 574 p.

Ostle, B. and R. W. Mensing. 1975. Statistics in research. Third edition, The Iowa State University Press, Ames, Iowa. 596 p.

Pielou, E. C. 1977. Mathematical ecology. John Wiley & Sons, New York. 305 p.

Pinkham, C. F. A. and J. G. Pearson. 1974. A new measure of biotic similarity between samples and its applications with a cluster analysis program. Department of the Army, Edgewood Arsenal Technical Report No. EB-TR-74062. 15 p.

Sokal, R. R. and P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H. Freeman and Company, San Francisco, CA. 359 p.

Williams, W. T. 1971. Principles of clustering. Ann. Rev. Ecol. Syst. 2:303-26.